

# Nonparametric Bayes modeling with sample survey weights

BY T. KUNIHAMA

*Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.*  
tsuyoshi.kunihama@duke.edu

A. H. HERRING

*Department of Biostatistics and Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.*  
aherring@bios.unc.edu

C. T. HALPERN

*Department of Maternal and Child Health and Carolina Population Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.*  
carolyn.halpern@unc.edu

AND D. B. DUNSON

*Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.*  
dunson@duke.edu

## SUMMARY

In population studies, it is standard to sample data via designs in which the population is divided into strata, with the different strata assigned different probabilities of inclusion. Although there have been some proposals for including sample survey weights into Bayesian analyses, existing methods require complex models or ignore the stratified design underlying the survey weights. We propose a simple approach based on modeling the distribution of the selected sample as a mixture, with the mixture weights appropriately adjusted, while accounting for uncertainty in the adjustment. We focus for simplicity on Dirichlet process mixtures but the proposed approach can be applied more broadly. We sketch a simple Markov chain Monte Carlo algorithm for computation, and assess the approach via simulations and an application.

*Some key words:* Biased sampling; Dirichlet process; Mixture model; Stratified sampling; Survey data.

## 1. INTRODUCTION

In sample surveys, it is routine to conduct stratified sampling designs to ensure that a broad variety of groups are adequately represented in the sample. In particular, the population is divided into mutually exclusive strata having different probabilities of inclusion. Analyzing data from such designs is challenging, since the collected sample is not representative of the overall population. To correct for discrepancies in the statistical analysis, survey weights are constructed. However, it is unclear how to appropriately include these weights, particularly in Bayesian analyses.

Little (2004) and Gelman (2007) clarify the importance of including survey weights into model-based analyses. Zheng & Little (2003, 2005) propose a nonparametric spline model and Chen et al. (2010) extend the framework for binary variables. An unpublished 2014 technical report by Y. Si, N. Pillai and A. Gelman propose a nonparametric model in which the survey weights are linked with a response through a Gaussian process regression. These approaches can flexibly connect the survey weights with the response. However, they require additional modeling of survey weights for non-sampled subjects in the population, leading to highly complex models.

In this article, we propose a simple approach in which we apply standard mixture models for the selected sample, and then adjust the mixture weights based on the survey weights. We allow probabilistic uncertainty in this adjustment in a Bayesian manner. Posterior computation relies on a simple modification to add an additional step to Markov chain Monte Carlo algorithms for mixture models.

## 2. MIXTURE MODELS WITH SURVEY WEIGHTS

### 2.1. Adjusted density estimates

Let  $y_1, \dots, y_N$  denote independently and identically distributed observations from a density  $f_0$  with  $y_i \in \mathcal{R}$  for  $i \in D = \{1, \dots, N\}$ . From this initial population,  $n$  subjects are sampled, with  $w_i = c/\pi_i$  the survey weight for subject  $i$ ,  $c$  a positive constant, and  $\pi_i$  the inclusion probability for  $i \in D$ . We assume  $D$  can be divided into mutually exclusive subpopulations  $D_1, \dots, D_M$ , with  $\{y_i, i \in D_m\}$  independently and identically distributed from density  $f_m$ , for  $m = 1, \dots, M$ . Then,  $f_0$  can be expressed as

$$f_0(y) = \sum_{m=1}^M \nu_m f_m(y), \quad (1)$$

where  $\nu_m \geq 0$  and  $\sum_{m=1}^M \nu_m = 1$ . By applying kernel density estimation to each  $f_m$  in (1), Buskirk (1998) and Bellhouse & Stafford (1999) propose an adjusted density estimate,

$$\hat{f}_0(y) = \sum_{i \in S} \frac{\tilde{w}_i}{b} \mathcal{K} \left( \frac{y - y_i}{b} \right), \quad (2)$$

where  $S \subset D$  are the selected subjects in the survey,  $\tilde{w}_i = w_i / \sum_{j \in S} w_j$ ,  $\mathcal{K}$  is a kernel function and  $b > 0$ . Estimator (2) adjusts for bias in the usual kernel estimator applied to sample  $S$  by modifying the weight for the  $i$ th subject from  $1/n$  to  $\tilde{w}_i$ . This adjustment leads to consistency under some conditions (Buskirk & Lohr (2005)).

### 2.2. Bayesian adjustments with uncertainty

Section 2.1 focuses on univariate continuous variables, while our goal is to develop a general approach for adjusting posterior distributions to take into account sample survey weights. Let  $y \in \mathcal{Y}$  denote a random variable, with  $\mathcal{Y}$  a Polish space that may correspond to a  $p$ -dimensional Euclidean space, a discrete space, a mixed continuous and discrete space, a non-Euclidean Riemannian manifold, such as a sphere, and other cases. Extending (1) to general spaces, we let  $f_0(\cdot)$  and  $f_m(\cdot)$ , for  $m = 1, \dots, M$ , denote densities on  $\mathcal{Y}$  with respect to a dominating measure

$\mu$ . The density in the  $m$ th subpopulation is expressed as a mixture,

$$f_m(y) = \sum_{h=1}^H \nu_{mh} f(y | \theta_h), \quad (3)$$

where  $\nu_{mh} \geq 0$ ,  $\sum_{h=1}^H \nu_{mh} = 1$  and  $\theta_h$  are parameters characterizing the  $h$ th mixture component. Then,  $f_0$  can be approximately expressed as a mixture having the same kernels as in (3) but with adjusted weights as in (2).

**THEOREM 1.** *Let  $s_i \in \{1, \dots, H\}$  denote the mixture index for subject  $i$  for  $i \in S$ . Let  $S_h = \{i : s_i = h, i \in S\}$ , for  $h = 1, \dots, H$ . Then, for large  $N$  and  $n$ ,*

$$f_0(y) \approx \sum_{h=1}^H \frac{\sum_{i \in S_h} w_i / c}{N} f(y | \theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y | \theta_{s_i}). \quad (4)$$

*Proof.* Letting  $N_m$  be the number of subjects in  $D_m$ ,  $N_m/N \rightarrow \nu_m$  by the law of large numbers. Letting  $w_m^*$  and  $\pi_m^*$  denote the survey weight and inclusion probability for the  $m$ th subpopulation,  $w_i = w_m^*$  and  $\pi_i = \pi_m^*$  for  $i \in D_m$ . From (1) and (3),  $f_0$  can be expressed as

$$\begin{aligned} f_0(y) &= \sum_{m=1}^M \nu_m f_m(y) \approx \sum_{m=1}^M \frac{N_m}{N} f_m(y) = \sum_{h=1}^H \sum_{m=1}^M \frac{N_m \nu_{mh}}{N} f(y | \theta_h) \\ &\approx \sum_{h=1}^H \frac{\sum_{i \in S_h} w_i / c}{N} f(y | \theta_h) \approx \sum_{i \in S} \tilde{w}_i f(y | \theta_{s_i}). \end{aligned} \quad (5)$$

The first approximation in (5) can be induced by  $N_m \approx w_m^* n_m / c$  and

$$\nu_{mh} \approx \frac{\sum_{i \in S} 1(i \in D_{mh})}{n_m},$$

for large  $N_m$  and  $n_m$ , where  $D_{mh}$  is a subset of  $D_m$  with  $s_i = h$ . The second approximation in (5) is based on  $c \approx \sum_{i \in S} w_i / N$ , which is derived by summation of  $N_m \approx w_m^* n_m / c$  over  $m$ .  $\square$

Under random designs with  $w_i \propto c$ ,  $f_0$  can be approximated by

$$f_R(y) = \sum_{h=1}^H \frac{\sum_{i \in D} 1(i \in S_h)}{n} f(y | \theta_h) = \sum_{i \in S} \frac{1}{n} f(y | \theta_{s_i}). \quad (6)$$

Comparing the last terms in (4) and (6), we can interpret that the bias can be adjusted by shifting the weight for the  $i$ th sampled subject from  $1/n$  to  $\tilde{w}_i$  as in (2).

We propose a simple Bayesian adjustment method using the second term in (4). We consider a standard Bayesian mixture model,

$$f_B(y) = \sum_{h=1}^H \lambda_h f(y | \theta_h), \quad \lambda \sim \pi(\lambda), \quad \theta_h \sim \pi(\theta_h), \quad (7)$$

where  $\lambda = (\lambda_1, \dots, \lambda_H)^T$  with  $\lambda_h \geq 0$  and  $\sum_{h=1}^H \lambda_h = 1$ , and  $\pi(\lambda)$  and  $\pi(\theta_h)$  are priors for  $\lambda$  and  $\theta_h$ . For example, using a truncated stick-breaking process (Ishwaran & James (2001)), we let  $\lambda_h = V_h \prod_{l < h} (1 - V_l)$ ,  $V_h \sim \text{Be}(1, \alpha)$  for  $h = 1, \dots, H - 1$  with  $V_H = 1$ . However, our focus is not on the specific mixture model and prior but on the adjustment for sampling bias, and alternative priors can be used without complication.

Comparing the second terms in (4) and (6), the difference is in the mixture weights. The expression in (4) can be interpreted as implying that  $\sum_{i \in S_h} w_i/c$  subjects are generated from the  $h$ th mixture component in the population. Updating the prior  $\tilde{\lambda} \sim \text{Dir}(a_1, \dots, a_H)$  with this information, we obtain the following conditional posterior distribution for the adjusted weights  $\tilde{\lambda} = (\tilde{\lambda}_1, \dots, \tilde{\lambda}_H)^T$ ,

$$\tilde{\lambda} \sim \text{Dir} \left( a_1 + \frac{1}{\tilde{c}} \sum_{i:s_i=1} w_i, \dots, a_H + \frac{1}{\tilde{c}} \sum_{i:s_i=H} w_i \right), \quad (8)$$

where  $\tilde{c} = \sum_{i \in S} w_i/N \approx c$ . Expression (8) takes into account uncertainty in the adjusted weights in mixture component allocation. Even as the population size  $N$  becomes large, there may be certain mixture components that are not represented in the selected sample, leading to substantial uncertainty in the adjustment. Posterior computation is straightforward: we simply apply any existing Markov chain Monte Carlo algorithm for mixture models to the selected sample, add sampling step (8) for generating the adjusted weights  $\tilde{\lambda}$ , and apply this adjustment to each step of the sampling algorithm to obtain samples from an adjusted posterior for the population density  $f_0(y)$ . As a default, we set  $a_h = a$  for  $h = 1, \dots, H$ , with prior sample size  $Ha \sim 1 - 2\%$  of population size  $N$ .

### 3. SIMULATION STUDY

We illustrate performance of the proposed approach and compare to competitors. We consider three cases in which a population with  $N = 1,000,000$  consists of three subpopulations having  $N_1 = 650,000$ ,  $N_2 = 300,000$  and  $N_3 = 50,000$  with  $\nu_m = N_m/N$ . From each stratum, we randomly generate  $n_m = 500$  subjects and construct survey weights by  $w_i = N_m/n_m$  for  $i \in D_m$  for  $m = 1, 2, 3$ . As competitors, we employ three model-based Bayesian methods. First, we consider a model-based Horvitz-Thompson estimator (Horvitz & Thompson (1952); Little (2004)),  $y_i = \beta\pi_i + \varepsilon_i$ ,  $\varepsilon_i \sim N(0, \pi_i^2\sigma^2)$  where  $\pi_i = 1/w_i$ . Second, we consider a polynomial regression with random effects,  $y_i = \beta_0 + \beta_1\pi_i + \beta_2\pi_i^2 + \gamma_{[i]} + \varepsilon_i$ ,  $\varepsilon_i \sim N(0, \pi_i^2\sigma^2)$ ,  $\gamma_m \sim N(0, \tau^2)$  where  $\gamma_{[i]}$  denotes a random effect for the subpopulation to which the  $i$ th subject belongs. This can be induced by the spline model of Zheng & Little (2003). Also, we apply the Gaussian process regression model from a 2014 technical report by Y. Si, N. Pillai and A. Gelman,  $y_i = \mu(x_{[i]}) + \varepsilon_i$ ,  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $\mu(x) \sim \text{GP}(\beta x, C)$ ,  $C\{\mu(x_m), \mu(x_{m'})\} = \text{cov}\{\mu(x_m), \mu(x_{m'})\} = \tau^2 \exp(-\kappa|x_m - x_{m'}|)$  where  $x_m = \log(w_m^*)$  and  $x_{[i]}$  denotes the log weight for the stratum for the  $i$ th subject. We also apply Dirichlet process mixtures without weight adjustment.

In the first case, we assume  $f_1(y) = f_N(y|2, 0.6)$ ,  $f_2(y) = f_N(y|0, 0.4)$  and  $f_3(y) = f_N(y|-2, 0.3)$  in (1) where  $f_N(y|a, b)$  denotes a normal density with mean  $a$  and standard deviation  $b$ . For the proposed method, we use the Dirichlet process mixture of normals,  $f_B(y) = \sum_{h=1}^H \lambda_h f_N(y|\mu_h, \tau_h)$  where  $\lambda_h = V_h(1 - V_l)$ ,  $V_h \sim \text{Be}(1, \alpha)$ ,  $V_H = 1$  with  $H = 20$ ,  $\alpha \sim \text{Ga}(0.25, 0.25)$ ,  $\mu_h \sim N(\bar{y}, s_y^2)$ ,  $\tau_h^2 \sim \text{Inverse-Gamma}(2, s_y^2/2)$  where  $\bar{y}$  and  $s_y^2$  are the sample mean and variance. As for the prior in the step (8), we set  $a_h = 1,000$  for each  $h$ . For competitors, we assume the following priors:  $\beta \sim N(0, s_y^2)$ ,  $\beta_j \sim N(0, s_y^2)$ ,  $\sigma^2 \sim \text{Inverse-Gamma}(2, s_y^2/2)$ ,  $\tau^2 \sim \text{Inverse-Gamma}(2, s_y^2/2)$  and  $\kappa \sim \text{Ga}(1, 2)$ . We draw 10,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. Rates of convergence and mixing were adequate. Figure 1 shows the estimation results for case 1. The Horvitz-Thompson estimator fails to capture the multimodality, while the

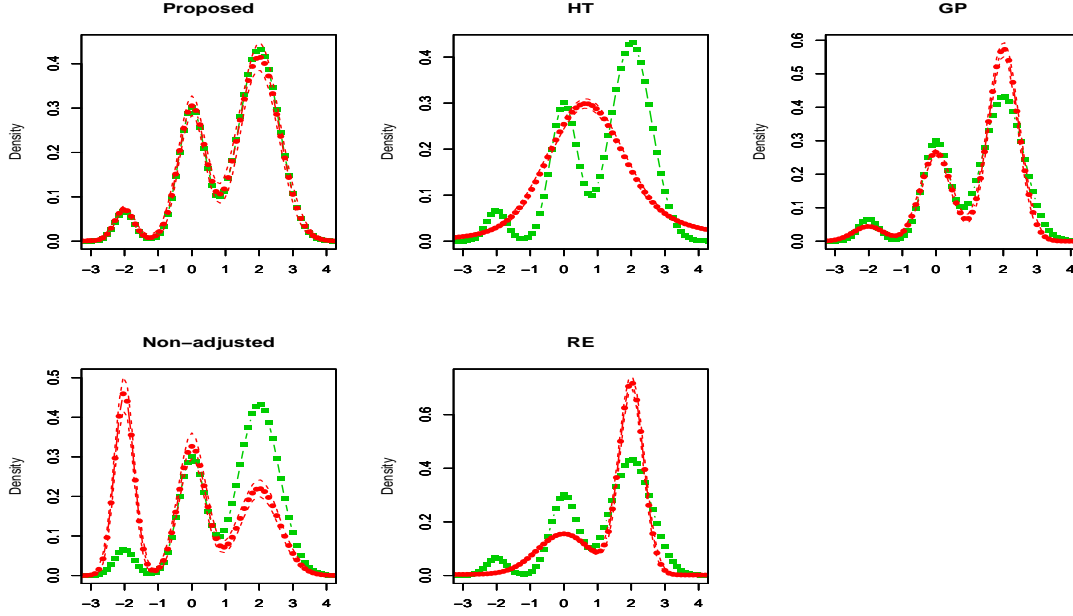


Fig. 1. Estimated densities in case 1. Green lines with squares are the true density, red lines with circles the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

non-adjusted estimator has considerable bias. The random effect model and Gaussian process have somewhat better performance, but clear bias remains. The proposed method accurately estimates the density, and 98% of true values are covered in the 95% credible intervals across 100 equally spaced grid points in  $[-6, 6]$ .

We also considered a more complex density for each stratum,  $f_1(y) = 0.2f_N(y | -2, 1) + 0.8f_N(y | 2, 0.8)$ ,  $f_2(y) = 0.4f_N(y | -2, 1) + 0.6f_N(y | 2, 0.8)$  and  $f_3(y) = 0.85f_N(y | -2, 1) + 0.15f_N(y | 2, 0.8)$ . The Markov chain Monte Carlo settings are the same as in case 1. Figure 2 reports the result for case 2. The Horvitz-Thompson estimator, random effect model and Gaussian process regression work poorly, missing the multimodal shape of the true density because they construct population densities relying on unimodal densities for subpopulations. The non-adjusted method capture the bimodality but with substantial bias. The proposed method approximates the density well, while covering 100% of true values in the 95% intervals.

We also consider a mixture of Poisson distributions,  $f_1(y) = 0.2\text{Poisson}(y | 15) + 0.8\text{Poisson}(y | 4)$ ,  $f_2(y) = 0.4\text{Poisson}(y | 15) + 0.6\text{Poisson}(y | 4)$  and  $f_3(y) = 0.85\text{Poisson}(y | 15) + 0.15\text{Poisson}(y | 4)$ . For the Dirichlet process mixtures, we apply the rounded kernel method in Canale & Dunson (2011) where latent continuous variables are modeled by (7) with the same Markov chain Monte Carlo settings as in case 1. Also, we apply the competitors to log transformed observations  $y_i^* = \log(y_i + 0.5)$  and estimate probabilities by  $\text{pr}(y_i = y) = \text{pr}\{\log(y) < y_i^* \leq \log(y + 1)\}$  for  $y = 0, 1, \dots, \infty$ . Figure 3 shows the result. We observe the proposed method obtains good approximation, while the competitors fail to capture the mode at 15. Also, 98% of the true values are covered in the 95% intervals in the support from 0 to 100.

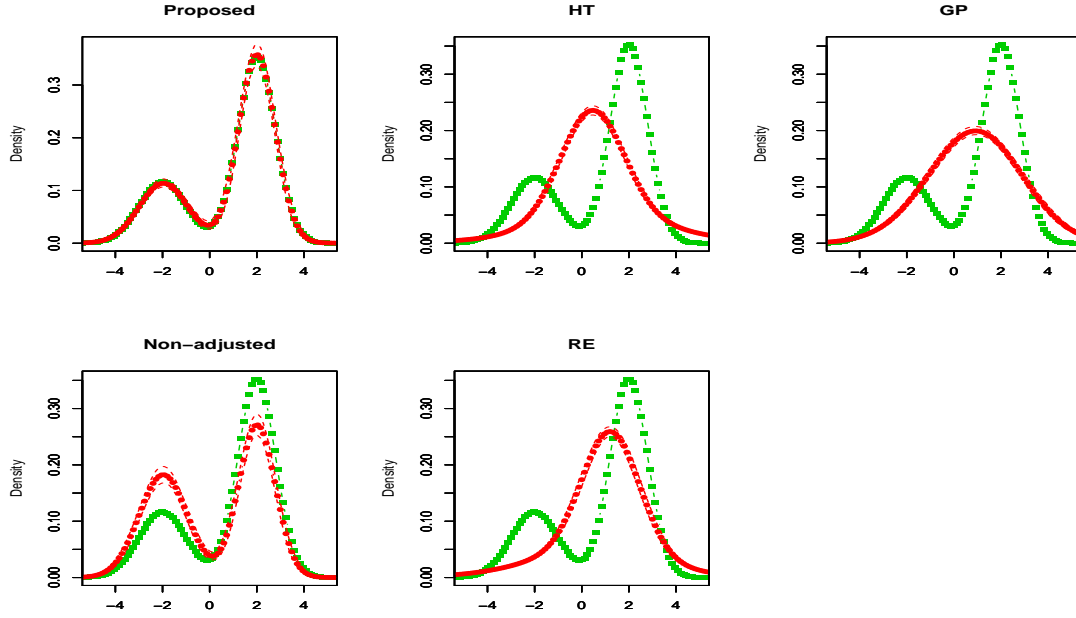


Fig. 2. Estimated densities in case 2. Green lines with squares are the true density, red lines with circles the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

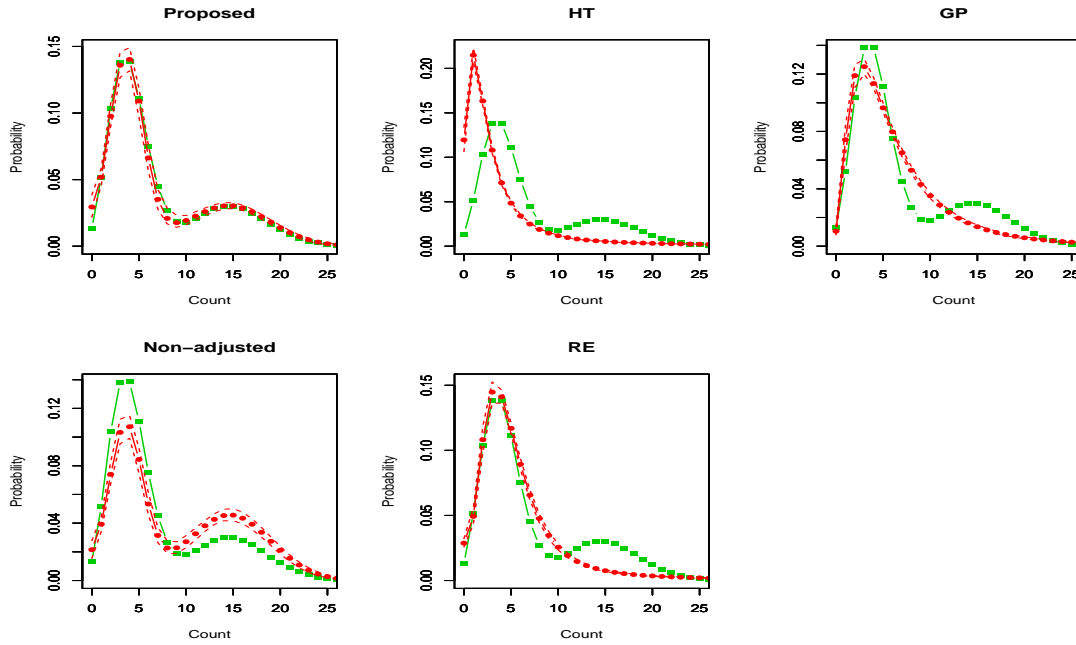


Fig. 3. Estimated probabilities in case 3. Green lines with squares are the true density, red lines with circles the posterior means and red dash lines 95% credible intervals. Proposed means the proposed method, Non-adjusted the Dirichlet process mixtures without weight adjustment, HT Horvitz-Thompson estimator, RE polynomial regression with random effects and GP Gaussian process regression.

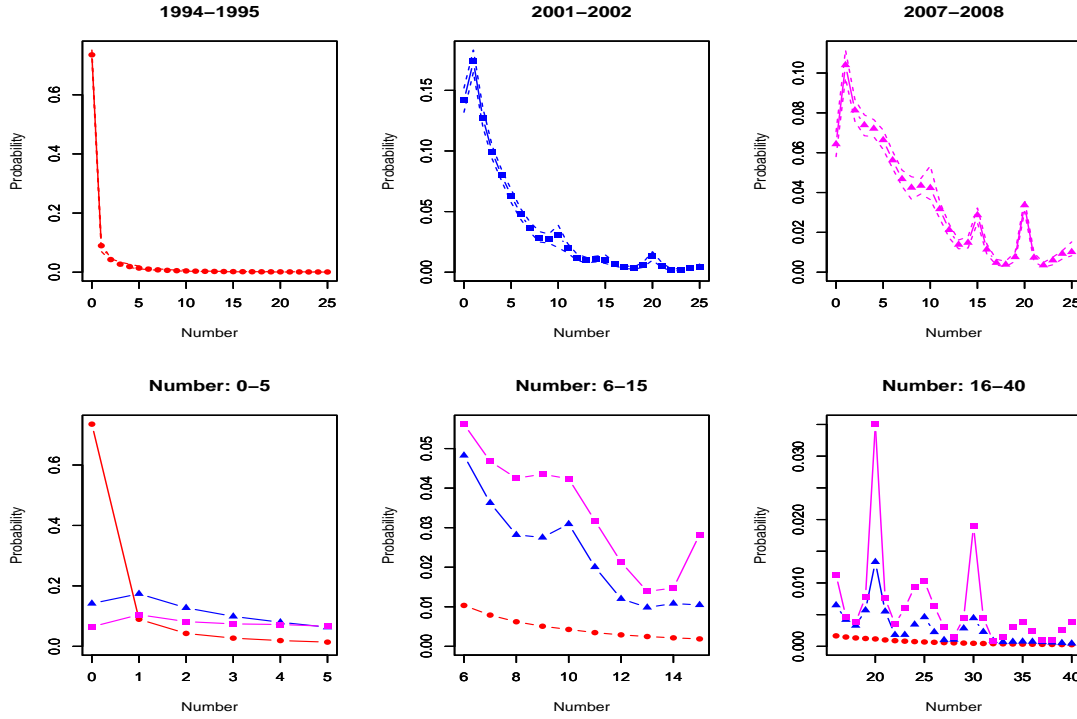


Fig. 4. Estimated probabilities of total numbers of sex partners. The first row shows estimated probabilities in 1994-1995 (left), 2001-2002 (middle) and 2007-2008 (right). Lines with symbols show posterior means and dash lines 95% credible intervals. The second row shows posterior means of 0-5 partners (left), 6-15 (middle) and 15-40 (left). Red lines with circle represent posterior means for 1994-1995, blue lines with triangles for 2001-2002 and purple lines with squares for 2007-2008.

To assess the impact of increasing the number of strata while decreasing within-strata sample size, we consider a case with  $M = 100$  in which  $N_m = 1000m$ ,  $n_m = 20$  for  $m = 1, \dots, 100$  with  $N = 5,050,000$  and  $n = 2,000$  and  $f_m(y) = f_N(y | -2, 0.3)$  for  $m = 1, \dots, 30$ ,  $f_m(y) = f_N(y | 0, 0.4)$  for  $m = 31, \dots, 70$  and  $f_m(y) = f_N(y | 2, 0.6)$  for  $m = 71, \dots, 100$ . We obtain a similar result to case 1 with the proposed method dominating competitors.

#### 4. APPLICATION TO ADOLESCENT BEHAVIOUR ANALYSIS

We apply the proposed method to the National Longitudinal Study of Adolescent Health. Our focus is on studying the total number of sex partners in adolescence. The target population is adolescents in grades 7-12 in the United States during the 1994-95 school year with  $N = 14,677,347$ . The full study design is described by Harris et al. (2009). The study drew supplemental samples, oversampling groups of particular interest based on ethnicity, genetic relatedness to siblings, adoption status, disability, and black adolescents with highly educated parents. We use three waves of surveys in which participants are in grades 7-12 (1994-1995), young adults age 18-26 (2001-2002) and adults age 24-32 (2007-2008). In each wave, numbers of observations are 6447, 4812 and 4819, respectively. We use the rounded kernel method with Dirichlet process mixtures as in the simulation. Since we expect high right skew in these data, we use log cut-points instead of non-negative integers, so that the Dirichlet process mixtures

can efficiently approximate such distributions. For the priors of the latent continuous variable, we use  $\mu_h \sim N(\tilde{y}, \tilde{s}_y^2)$ ,  $\tau_h^2 \sim \text{Inverse-Gamma}(2, \tilde{s}_y^2/200)$  where  $\tilde{y}$  and  $\tilde{s}_y^2$  are the sample mean and variance of  $\log(y_i + 0.5)$ . Also, we set  $a_h = 10,000$  for the Dirichlet prior in (8). We draw 20,000 samples after the initial 5,000 samples are discarded as a burn-in period and every 10th sample is saved. We observe that the sample paths were stable and the sample autocorrelations dropped smoothly.

Figure 4 shows the estimated probabilities for the three waves. 1994-1995 shows a high probability on zero with small values for positive counts. 2001-2002 expresses differences from 1994-1995 in that the probability on zero considerably decreases, while one shows the highest value and the tail gets heavy. The shape in 2007-2008 is similar to 2001-2002 in that both have highest probabilities at one and then steep declines. 2007-2008 shows a heavier tail with relatively high spikes at multiples of five. This is probably because people with many partners do not remember the exact numbers.

#### ACKNOWLEDGEMENT

This work was supported by Nakajima Foundation and grants from the National Institutes of Health. The data are from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. The data are available at <http://www.cpc.unc.edu/addhealth>.

#### REFERENCES

- BELLHOUSE, D. R. & STAFFORD, J. E. (1999). Density estimation from complex surveys. *Statistica Sinica* **9**, 407–424.
- BUSKIRK, T. D. (1998). Nonparametric density estimation using complex survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association*. Washington, DC.
- BUSKIRK, T. D. & LOHR, S. L. (2005). Asymptotic properties of kernel density estimation with complex survey data. *Journal of Statistical Planning and Inference* **128**, 165–190.
- CANALE, A. & DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association* **106**, 1528–1539.
- CHEN, Q., ELLIOTT, M. R. & LITTLE, R. J. A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology* **36**, 23–34.
- GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science* **22**, 153–164.
- HARRIS, K. M., HALPERN, C. T., WHITSEL, E., HUSSEY, J., TABOR, J., ENTZEL, P. & UDRY, J. R. (2009). The national longitudinal study of adolescent health: Research design [www document] URL: <http://www.cpc.unc.edu/projects/addhealth/design>.
- HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association* **47**, 663–685.
- ISHWARAN, H. & JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.
- LITTLE, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99**, 546–556.
- ZHENG, H. & LITTLE, R. J. A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics* **19**, 99–107.
- ZHENG, H. & LITTLE, R. J. A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics* **21**, 1–20.